Amplitude envelope kinematics of speech signal: parameter extraction and applications

He, Lei ; Dellwo, Volker

Abstract: In this paper, we model the amplitude envelope of the broadband speech signal as a kinematic system and calculate its basic parameters, including displacement, velocity and acceleration. Such system captures the smoothed amplitude fluctuation pattern over time, illustrating how energy is distributed across the signal. Although the pulmonic air pressure is the primary energy source of speech, the amplitude modulation pattern is largely determined by articulatory behaviours, especially mandible and lip movements. Therefore, there should be a correspondence between signal envelope kinematics and articulator kinematics. Previous research showed that a tremendous amount of speaker idiosyncrasies in articulation existed. Such idiosyncrasies should therefore be reflected in the envelope kinematics as well. From the signal envelope kinematics, it may be possible to infer individual articulatory behaviours. This is particularly useful for forensic phoneticians who usually have no access to articulatory data, and clinical speech pathologists who usually find it difficult to make articulatory measurement in clinical consultations. Also in this paper, we illustrate a correspondence between the amplitude envelope kinematics and the lower lip kinematics (X-ray pellet history data) of one speaker reading one sentence. For future research, more speakers are needed to record both speech and articulatory signals to build a statistical model between the kinematics data of both domains.

AMPLITUDE ENVELOPE KINEMATICS OF SPEECH SIGNAL: PARAMETER EXTRACTION AND APPLICATIONS

Lei He, Volker Dellwo

Phonetics Laboratory, Institute of Computational Linguistics, University of Zurich
{lei.he|volker.dellwo}@uzh.ch

Abstract: In this paper, we model the amplitude envelope of the broadband speech signal as a kinematic system and calculate its basic parameters, including displacement, velocity and acceleration. Such system captures the smoothed amplitude fluctuation pattern over time, illustrating how energy is distributed across the signal. Although the pulmonic air pressure is the primary energy source of speech, the amplitude modulation pattern is largely determined by articulatory behaviours, especially mandible and lip movements. Therefore, there should be a correspondence between signal envelope kinematics and articulator kinematics. Previous research showed that a tremendous amount of speaker idiosyncrasies in articulation existed. Such idiosyncrasies should therefore be reflected in the envelope kinematics as well. From the signal envelope kinematics, it may be possible to infer individual articulatory behaviours. This is particularly useful for forensic phoneticians who usually have no access to articulatory data, and clinical speech pathologists who usually find it difficult to make articulatory measurement in clinical consultations. Also in this paper, we illustrate a correspondence between the amplitude envelope kinematics and the lower lip kinematics (X-ray pellet history data) of one speaker reading one sentence. For future research, more speakers are needed to record both speech and articulatory signals to build a statistical model between the kinematics data of both domains.

1 Introduction

In this paper, we propose a method to calculate the basic kinematics parameters of the amplitude envelope (henceforth, ENV) of the broadband speech signal. These parameters include the ENV displacement, velocity and acceleration. We believe that such parameters can approximate measures of articulatory kinematics of mandibular (or lower lip) movements, especially when instruments of articulography are inaccessible. The paper is structured as follows: section 2 explains the rationale of calculating the ENV kinematics parameters; section 3 shows the signal processing details in the Praat environment [1]; section 4 illustrates a correspondence between the ENV kinematics and lower lip articulatory kinematics based on one sentence; and section 5 discusses potential applications of the ENV kinematics parameters.

2 Rationale

We model the ENV as a kinematic system because such system captures the smoothed amplitude fluctuation pattern over time, illustrating how energy is distributed across the signal. Although the pulmonic air pressure is the primary energy source of speech, the amplitude modulation pattern is largely determined by articulatory behaviours, especially mandible and lip movements, which underpin the overall rhythmicity of speech. There is already evidence that the ENV function and the mouth aperture function co-vary (see Figure 1 for an illustration) [2]: the peaks and troughs of both curves occur around similar time points. This suggests that the variation of the ENV magnitude is determined by the articulatory behaviours which stand in close relationship with the degree of mouth aperture, including the

mandibular movement and the lip movement. The vertical movement of the lower lip is largely congruent with the vertical jaw displacement in normal speech. Therefore, we can estimate the kinematic characteristics of the jaw or lower lip from the ENV.

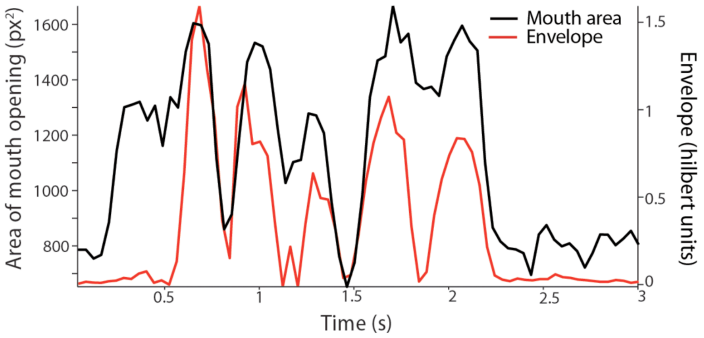

Figure 1 – An illustration of the covariation of both ENV function and mouth aperture function. The figure was originally published in [2] which is distributed under the terms of the Creative Commons Attribution License.

There have already been several studies which investigated articulation characteristics by measuring different aspects in the speech signal:

- [3, 4, 5] measured temporal variability as well as the ENV spectrum of the signal to distinguish different types of dysarthria;
- [6, 7] measured temporal variability of the speech signal to test speaker individuality resulted from idiosyncratic articulation;
- [8, 9, 10, 11] measured intensity variability of the speech signal to test speaker individuality resulted from idiosyncratic articulation.

To our knowledge, the method described here is the first attempt to estimate the kinematic properties of the jaw (or lower lip) from the ENV kinematics. Very similar to the approach described in this paper is [10, 11], where we investigated the intensity ramping patterns (i.e., the averaged speeds of intensity increases and decreases between alternating peaks and troughs), and discovered that the intensity decreases (corresponding to the closing gestures of the jaw) explained more between-speaker variability.

3 Signal processing

We use the pressure wave signal (Figure 2-*a*) of the sentence "Your good pants look great, however, your ripped pants look like a cheap version of K-mart special" spoken by one speaker to illustrate all signal processing details [(1)].

To obtain the ENV of the speech signal, we apply the Hilbert transform to construct the complex-valued analytic signal. The raw ENV is extracted by taking the complex modulus of the analytic signal. More details of this processing step are explained in [12]. However, for a broadband signal like this the extracted ENV contains a large amount of high-frequency noise [13]. Therefore, we low-pass filter the raw ENV to obtain a smoothed ENV function (cut-off frequency = 5 Hz, smoothing = 5 Hz). The 5 Hz cut-off is chosen because for clean speech, the ENV spectrum has a peak at ~5 Hz, reflecting the energy fluctuations associated with articulatory gestures corresponding to syllables [14]. The smoothed ENV itself (Figure 2-*b*) is

[(1)] The waveform of this sentence and its corresponding lower lip articulation signal (illustrated in Sec. 4) were downloaded from http://sail.usc.edu/~lgoldste/General_Phonetics/Week2/Gestures_new/Gestures.html (last accessed on 31.12.2016)

the displacement curve of the ENV kinematic system. We use $e(t)$ to notate this function. Moreover, we linearly normalise the smoothed ENV so that the highest amplitude is equal to unit 1, thus remove the artefact that speakers may be recorded with different gains. We refer to the unit of the ENV displacement as the normalised Hilbert unit (abbreviated as nHU). The script lines 5–21 in Figure 3 demonstrated the processing steps using Praat.

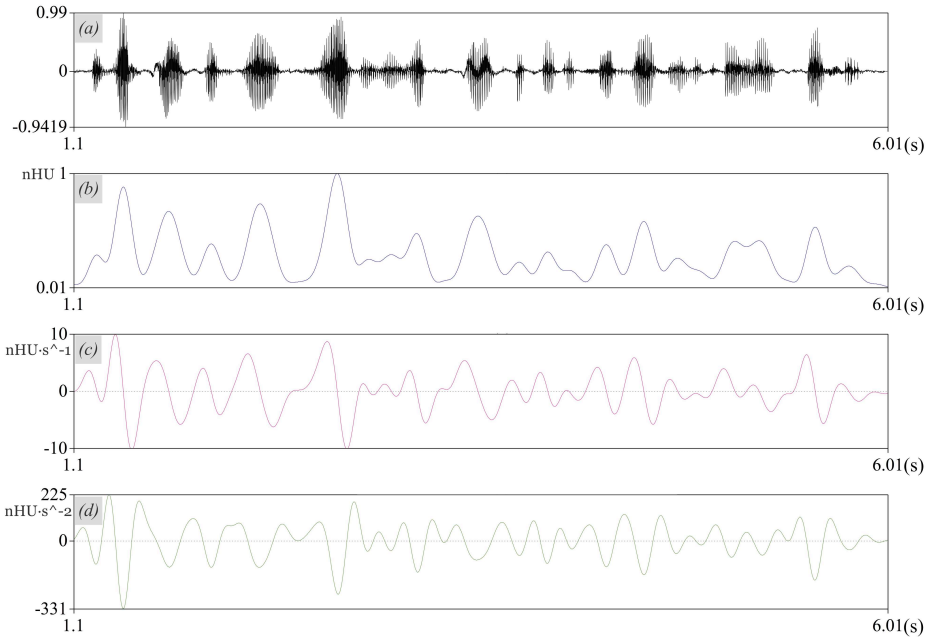

Figure 2 – The ENV kinematics parameters of a speech signal. Subplot (*a*) shows the original waveform; subplot (*b*) shows the ENV displacement curve $e(t)$; subplot (*c*) shows the ENV velocity curve $\dot{e}(t)$; and subplot (*d*) shows the ENV acceleration curve $\ddot{e}(t)$.

To derive the velocity curve of the ENV (notated as $\dot{e}(t)$), we simply need to take the first-order derivative of the ENV displacement curve $e(t)$. Since we are working with a digitised signal, the first-order discrete derivative can be calculated using the difference equation:

$$\dot{e}(t) \stackrel{\text{def}}{\equiv} \dot{e}[n] = \big(e[n+1] - e[n]\big)/T$$

where n = the index of sample numbers, and T = the sampling period of the signal. The result is illustrated in Figure 2-*c*. The unit for the ENV velocity is $\text{nHU}\cdot\text{s}^{-1}$.

To derive the acceleration curve of the ENV (notated as $\ddot{e}(t)$), we need to take the second-order derivative of the ENV displacement function $e(t)$, which is equivalent to the first-order derivative of the velocity function $\dot{e}(t)$. This process can also be approximated using the difference equation:

$$\ddot{e}(t) \stackrel{\text{def}}{\equiv} \ddot{e}[n] = \big(\dot{e}[n+1] - \dot{e}[n]\big)/T$$

where n = the index of sample numbers, and T = the sampling period. The derived ENV acceleration curve is illustrated in Figure 2-*d*. The unit for the ENV acceleration is $\text{nHU}\cdot\text{s}^{-2}$.

The script lines 22–27 in Figure 3 show the calculations of ENV velocity and acceleration curves in Praat. At this moment, all the basic kinematics parameters of ENV (displacement, velocity and acceleration) have been extracted [(2)].

envKinematics.txt — UNREGISTERED

```
# Sec. 0 # This script assumes that a sound signal is selected
# Sec. 1 # Get ID and name of selected sound
    sound = selected("Sound")
    name$ = selected$("Sound")
# Sec. 2 # Hilbert transform to obtain ENV
    # This step is based on the following publication:
    # He, Lei; Dellwo, Volker (2016) http://dx.doi.org/10.21437/Interspeech.2016-1447
    spectrum = To Spectrum: "no"
    Rename: "original"
    hilbertSpectrum = Copy: "hilbert"
    Formula: "if row=1 then Spectrum_original[2,col] else -Spectrum_original[1,col] fi"
    hilbertSound = To Sound
    envUnsmoothed = Formula: "sqrt(self^2 + Sound_'name$'[]^2)"
    Rename: "'name$'_ENV_Unsmoothed"
    removeObject: spectrum, hilbertSpectrum
# Sec. 3 # Get a smoothed ENV
    env = Filter (pass Hann band): 0, 5, 5
    Rename: "'name$'_ENV"
    removeObject: envUnsmoothed
# Sec. 4 # Scale peak of ENV
    Scale peak: 1
# Sec. 5 # Obtain the first order derivative of ENV – envelope velocity curve
    vel = Copy: "'name$'_VEL"
        Formula: "(self[col+1]-self)/dx"
# Sec. 6 # Obtain the second order derivative of ENV – envelope acceleration curve
    acc = Copy: "'name$'_ACC"
        Formula: "(self[col+1]-self)/dx"

                    ### END ###

```

Line 1, Column 1 — Tab Size: 4 — Plain Text

Figure 3 – A Praat code to extract the ENV kinematics parameters from a speech signal.

4 ENV kinematics vs. lower lip kinematics: an illustration

Now, we calculate the kinematics parameters of the lower lip articulatory trajectory (recorded using the X-ray microbeam technique) which was recorded contemporaneously with the speech signal illustrated in section 3. The trajectory itself is already a displacement function, from which we calculate the first- and second- order derivatives, which are the velocity and acceleration of the lower lip movement. The articulatory displacement, velocity and acceleration functions are notated as $a(t)$, $\dot{a}(t)$ and $\ddot{a}(t)$. They are also digitised signals, so their derivatives are calculated using difference equations the same way as we calculated the ENV kinematics:

$$\dot{a}(t) \stackrel{\text{def}}{=} \dot{a}[m] = \big(a[m+1] - a[m]\big)/T_{\text{x}}$$

$$\dot{a}(t) \stackrel{\text{def}}{=} \ddot{a}[m] = \big(\dot{a}[m+1] - \dot{a}[m]\big)/T_{\text{x}}$$

where m = the index of X-ray sample numbers, and T_{x} = the sampling period of the X-ray signal.

[(2)] Differentiation is much easier using general-purpose programming languages such as MATLAB and Scilab with the function `diff(X,n)`, where `X` refers to the vector to be differentiated, and `n` refers to the order of differentiation.

We juxtopose the kinematics functions of both ENV and the lower lip together to illustrate their correspondences. Figures 4 – 6 illustrate the comparisons between $e(t)$ vs. $a(t)$, $\dot{e}(t)$ vs. $\dot{a}(t)$, and $\ddot{e}(t)$ vs. $\ddot{a}(t)$ respectively.

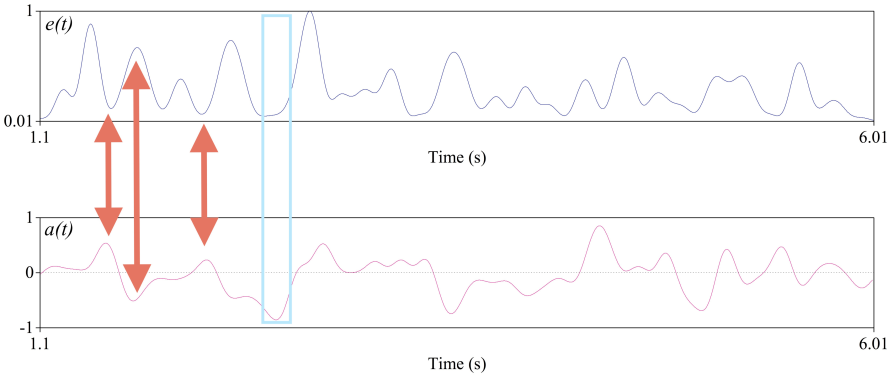

Figure 4 – The correspondence between the ENV displacement $e(t)$ and lower lip displacement $a(t)$.

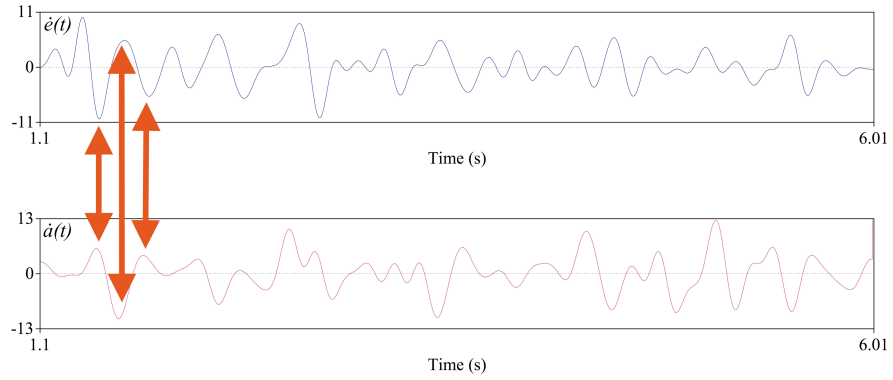

Figure 5 – The correspondence between the ENV velocity $\dot{e}(t)$ and lower lip displacement $\dot{a}(t)$.

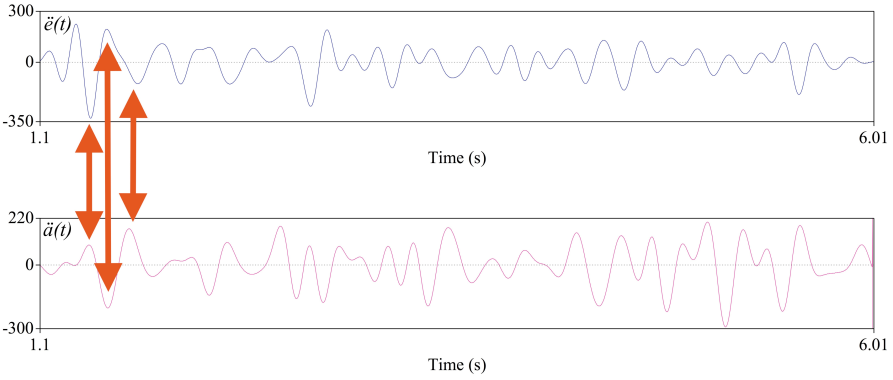

Figure 6 – The correspondence between the ENV velocity $\ddot{e}(t)$ and lower lip displacement $\ddot{a}(t)$.

From these graphs, we can see that a correspondence exists between the kinematics of both ENV and lower lip movements. Where there is a trough for the lip position, a peak in the ENV is usually observed. Such is also the case for velocity and acceleration for both ENV and lower lip movements. An exception to this pattern is also highlighted in the rectangular box in Figure 4, where an ENV trough is associated with a trough of lower lip position. That is because the speaker paused a little bit after "Your good pants look great". The position of the lower lip still remained in a low position, but there is little energy in the acoustic signal.

5 Outlooks

For future research, more speakers are needed to record both speech and articulatory signals to build a mathematical model between the kinematics of both domains. With a well-constructed model, we envision the applications of ENV kinematics in the following areas:

• *Forensic phonetics* Speakers possess a tremendous amount of individuality in all major aspects of speech: source signal, vocal tract resonance, and articulation [15]. It is possible to measure the first two aspects directly from the speech signal, but articulation can only be estimated indirectly. If the relationship between the ENV kinematics and articulator kinematics can be quantitatively and reliably established, it should be possible to estimate the articulatory behaviours from the ENV kinematics parameters. This should be particularly helpful to forensic experts when only video footages of a criminal are available. From such footages, measuring the kinematics of the articulators should be straightforward. During the police interrogations, obtaining the speech samples of the suspect is also easy. From the speech samples, it is possible to calculate the ENV kinematics. Then based on such measures, forensic experts may be able to estimate how likely the suspect and the criminal are the same person, given that a mathematical model between ENV and articulator kinematics is established.

• *Speech pathology and phoniatrics* One type of the pathological speech conditions are the motor speech disorders, including the dysarthria, apraxia and developmental verbal dyspraxia. Although the causes of these conditions are different, the patients all manifest different forms of abnormalities in articulation. Direct articulatory measurements are informative to the diagnosis of such abnormalities (e.g., [16, 17]). However, instrumental articulatory measurements are time-consuming, making it difficult to be implemented in clinical consultations. With more knowledge about the relationship between the two domains of kinematics, it might be possible in the future for speech pathologists and phoniatricians to make initial diagnosis using the ENV kinematics parameters automatically obtained from clinical recordings, and then decide whether further examinations are needed.

For both outlooks, there is a high demand for further in-depth research to obtain more knowledge between the two domains of kinematics.

References

[1] BOERSMA, P. and WEENINK, D.: Praat: doing phonetics by computer (version 6.0.19 for Macintosh), downloaded from www.praat.org, 2016.

[2] CHANDRASEKARAN, C., TRUBANOVA, A., STILLITTANO, S., CAPLIER, A. and GHAZANFAR, A. A.: The natural statistics of audiovisual speech. PLoS Computational Biology 5: e1000436, 2009.

[3] LISS, J. M., WHITE, L., MATTYS, S., LANSFORD, K., LOTTO, A. J., SPITZER, S. M. and CAVINESS, J. N.: Quantifying speech rhythm abnormalities in the dysarthrias. Journal of Speech Language and Hearing Research 52, pp. 1334 – 1352, 2009.

[4] WHITE, L., LISS, J. and DELLWO, V.: Assessment of rhythm. In: Lowit, A. & Kent, R.

D. (Eds.): Assessment of Motor Speech Disorders. San Diego, USA: Plural Publishing, pp. 312 – 352, 2010.

[5] LISS, J. M., LEGENDRE, S. and LOTTO, A. J.: Discriminating dysarthria type from envelope modulation spectra. Journal of Speech Language and Hearing Research 53, pp. 1246 – 1255, 2010.

[6] DELLWO, V., LEEMANN, A. and KOLLY, M.-J.: Rhythmic variability between speakers: articulatory, prosodic, and linguistic factors. Journal of the Acoustical Society of America 137, pp. 1513 – 1528, 2015.

[7] LEEMANN, A., KOLLY, M.-J. and DELLWO, V.: Speaker-individuality in suprasegmental temporal features: implications for forensic voice comparison. Forensic Science International 238, pp. 59 – 67, 2014.

[8] HE, L. and DELLWO, V.: The role of syllable intensity in between-speaker rhythmic variability. International Journal of Speech, Language and the Law 23, pp. 243 – 273, 2016.

[9] HE, L., GLAVITSCH, U. and DELLWO, V.: Comparisons of speaker recognition strengths using suprasegmental duration and intensity variability: an artificial neural networks approach. In: Proceedings of the 18th International Congress of Phonetic Sciences (ICPhS), Glasgow, UK, Paper #0395, 2015.

[10] HE, L., GLAVITSCH, U. and DELLWO, V.: Inter-speaker variability in intensity dynamics. Talk given at the 24th Annual Conference of the International Association for Forensic Phonetics and Acoustics (IAFPA), 8th – 10th July 2015, Leiden, the Netherlands.

[11] HE, L. and DELLWO, V.: Temporal organization of the speech signal: closing gestures of jaw movements are more speaker-specific than opening gestures. Paper under review.

[12] HE, L. and DELLWO, V.: A Praat-based algorithm to extract the amplitude envelope and temporal fine structure using the Hilbert transform. In: Proceedings of INTERSPEECH 2016, San Francisco, USA, pp. 530 – 534, 2016.

[13] SØNDERGAARD, P. L., DECORSIÈRE, R. and DAU, T.: On the relationship between multi-channel envelope and temporal fine structures. In: Dau, T., Jepsen, M. L., Poulsen, T. & Dalsgaard, J. C. (Eds.): Speech Perception and Auditory Disorders. Ballerup, Denmark: Danavox Jubilee Foundation, pp. 363 – 370, 2011.

[14] GREENBERG, S., CARVEY, H., HITCHCOCK, L. and CHANG, S.: Temporal properties of spontaneous speech – a syllable-centric perspective. Journal of Phonetics 31, pp. 465 – 485, 2003.

[15] DELLWO, V., HUCKVALE, M. and ASHBY, M.: How is individuality expressed in voice? An introduction to speech production and description for speaker classification. In: Müller, C (Ed.): Speaker Classification I: Fundamentals, Features and Methods. Berlin and Heidelberg, Germany: Springer, pp. 1 – 20, 2007.

[16] ACKERMANN, H., GRÖNE, B. F., HOCH, G. and SCHÖNLE, P. W.: Speech freezing in Parkinson's disease: a kinematic analysis of orofacial movements by means of electromagnetic articulography. Folia Phoniatrica et Logopaedica 45, pp. 84 – 89, 1993.

[17] BARTLE, C. J., GOOZÉE, J. V., SCOTT, D., MURDOCH, B. E. and KURUVILLA, M.: EMA assessment of tongue-jaw coordination during speech in dysarthria following traumatic brain injury. Brain Injury 20, pp. 529 – 545, 2006.